Lyman-α forest in three dimensions: Computation issues

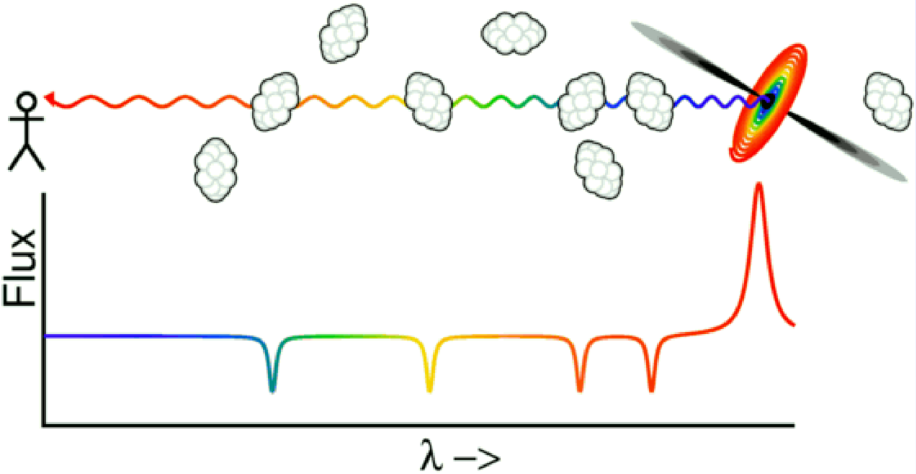
Anže Slosar & Nishikanta Khandai
(BNL)

ANL, SciDac Meeting, 10/18/2012

Introduction

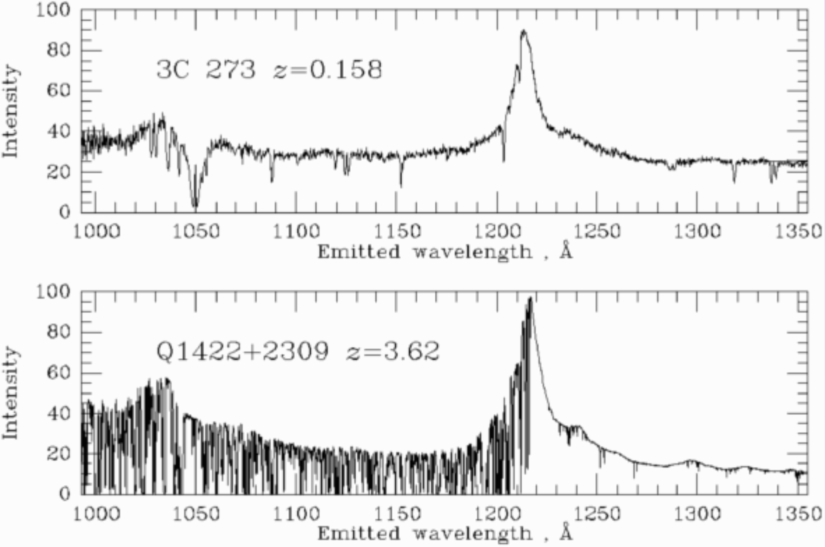
- Lyman-α forest is emerging as a 3D tracer of cosmic structure
- It presents serious computational issues:
  - Strong coupling of small scales to large scales, both in data analysis and theory.
  - In data analysis: small scale systematics can affect the large scale measurements of 2-point function
  - You want to be more clever than simply averging small scales, but the number of pixels is humongous 10^8
  - In theory: small scale fluctuations affect large scale linear bias parameters
- Talk plan:
  - Introduction to Lyman-α forest
  - Data analysis of BOSS data
  - Simulations and theoretical issues

Lyα forest



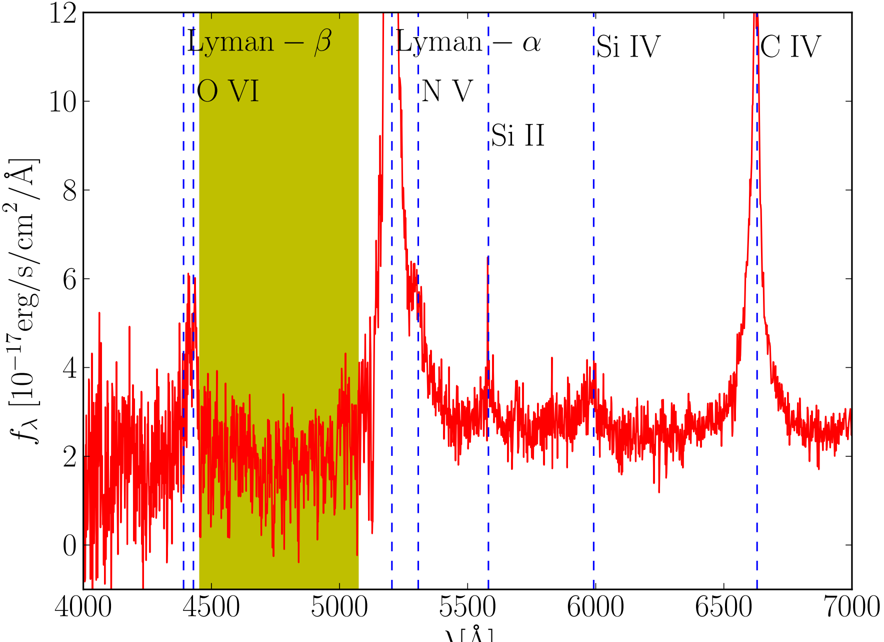
Neutral hydrogen absorbs light from distant quasars blue-ward of Lyα emission.

Lyα forest

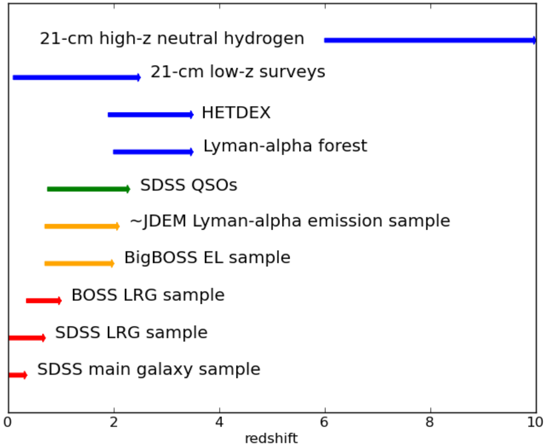


Neutral hydrogen absorbs light from distant quasars blue-ward of Lyα emission.

BOSS spectra



Measuring Density fields



- Lyman-α forest pretty unique in probing redhift 2-3 universe
- Volume probed is very, very large
- Systematics very different to galaxy surveys
- At $z < 2$ limited by forest moving into UV
- At $z > 3.5$ limited by faintness and number-density of quasars

Data reduction

- Data is big: Final survey will have some 150,000 quasars: each forest is only around 500 pixels, but to understand systematics you want to analyze entire quasars, so some 1500 pixels per quasar
- Ideally want to do analysis with two-point measurements sliced as much as possible: we used 3 redhift bins, 18 separation (perpendicular distance) bins and 28 $\Delta \log \lambda$ bins (parallel distance): 1512 measurements: barely enough to resolve BAO, ideally one would have more 5000 measurements.
- We used optimal estimator with *per-quasar* inverse covariance weighting: impossible to do at full resolution, so we compressed the data $\times 4$.
- Good point: all tasks are trivially parallelizable

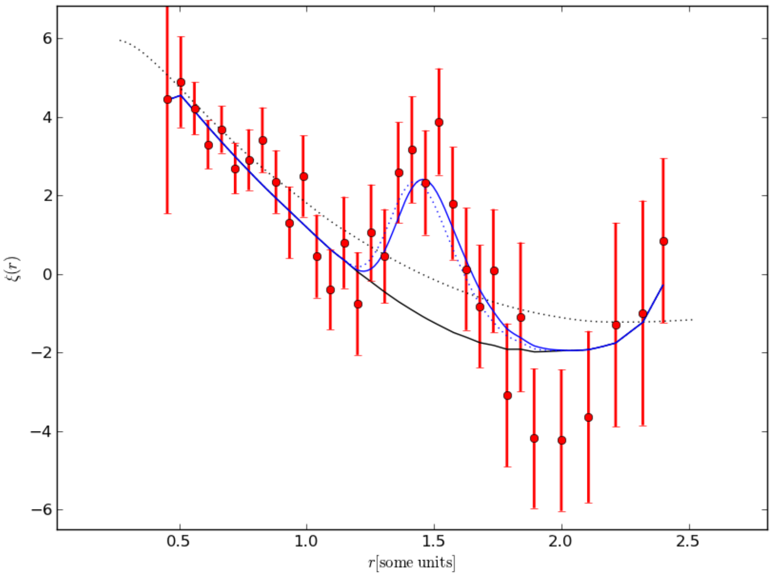
Quadratic estimator

- We're performing calculations of the kind

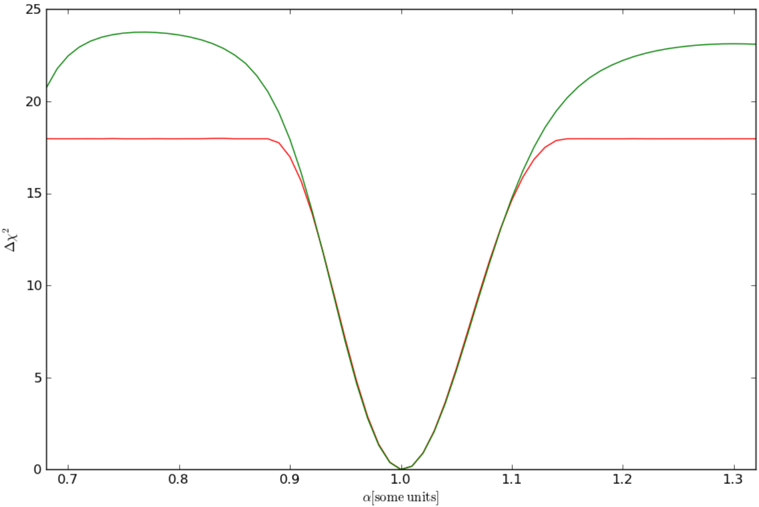
$$\begin{aligned} E_i &= \mathrm{Tr}(\mathrm{d}_1^T \mathrm{C}_1^{-1} \mathrm{C}_{,i} \mathrm{C}_2^{-1} \mathrm{d}_2) && (1) \\ F_{i,j} &= \frac{1}{2} \mathrm{Tr}(\mathrm{C}_1^{-1} \mathrm{C}_{,i} \mathrm{C}_2^{-1} \mathrm{C}_{,j}^T). && (2) \end{aligned}$$

- Common sense is that if you can calculate $\mathrm{C}^{-1}\mathrm{d}$ you win, but here, this is actually computationally fairy trivial. Typical size ~ 500 elements.
- The big problem is the Fisher matrix: $1512^2/2$ matrix multiplications for *each quasar pair*.
- We calculate $\mathrm{C}^{-1}\mathrm{d}$ and reduce pixel size after that. Survey doable at ×4 and ×3 compression, very hard lower compressions
- If measuring correlation function $C_{,i}$ is sparse.
- At ×1 compression, the sparse routines are considerably faster, at ×4 within 10% of dense matrices.

...and it kinda works



...and it kinda works



Improvements

- Even leaving the current technique unchanged, significant improvements can be gained from **GPU** utilization.
- One can fit 2000 500×500 matrices in 2Gb and GPUs should allow approximately $100\times$ speed-up on such problems
- This would allow one to do BOSS with no compression.
- Better probably to improve technique: high compression for widely separated pairs, no compression for closely separated pairs.
- Maybe do a FT-like transform first?

Improvements 2

- How to go beyond independent quasars approximation?
- The full problem is unfeasible
- Correlations beyond closest pairs small so some perturbation scheme should work.
- Most such schemes still require one to multiply $N_{\rm tot}$ sized matrices, which is likely to be prohibitively expensive.
- A good approach would be hierarchical smoothing: do low-k modes on smoothed full field, high-k modes on independent sub-volumes approximation.

Simulations of the Lyα Forest

Table: Simulation Parameters

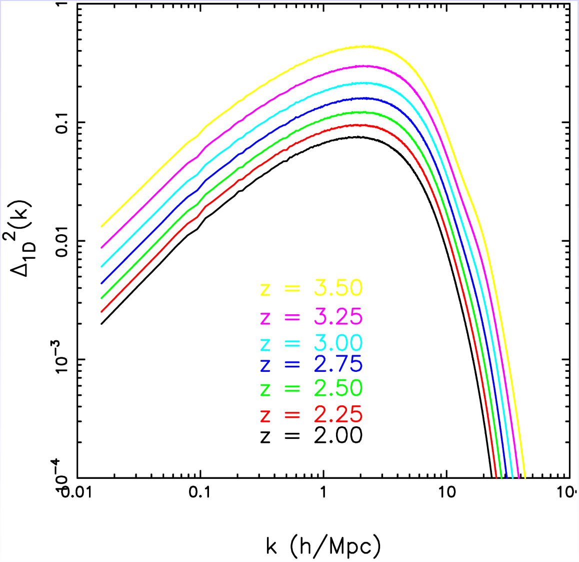
$L_{\rm box}$ $(h^{-1}{\rm Mpc})$	$N_{\rm part}$	$m_{\rm DM}$ $(h^{-1}M_{\odot})$	$m_{\rm gas}$ $(h^{-1}M_{\odot})$	ϵ $(h^{-1}{\rm kpc})$	z_f
400	2×4096^3	5.9×10^7	1.18×10^7	3.25	2.0

- Gadget3: DM, Gas, Star
- Cosmology: WMAP7
- Spectra created from gas properties, e.g. T, ρ, ε etc.

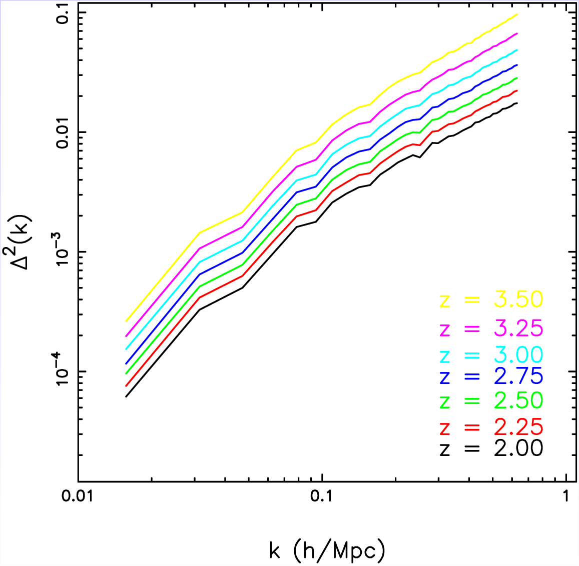
Figure: Kraken. University of Tennessee. 112896 cores, 147 Tb RAM

MassiveBlack : 6Tb/snapshot, 37 Snapshots, 98304 cores, $\sim 19 \times 10^6$ SUs.

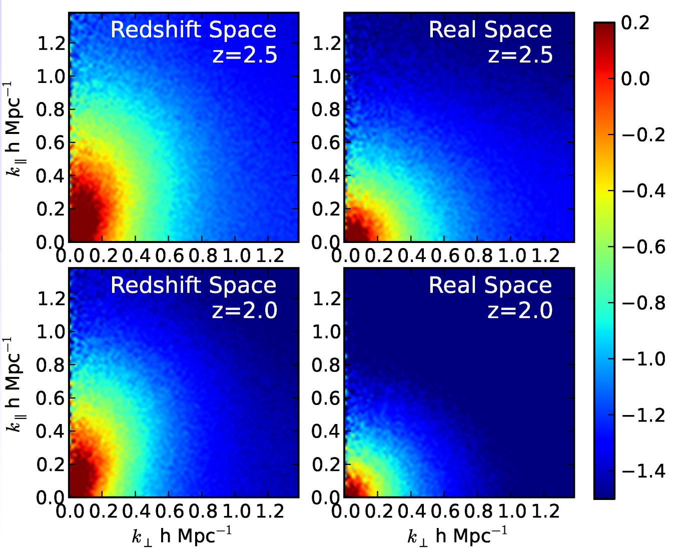
The 1D Lyα Forest Flux Power Spectrum



The 3D Lyα Forest Flux Power Spectrum



Redshift-space Distortions of the Lyα Forest Flux Power Spectrum



The bias model

- On large scales we relate $\delta_F(\mathbf{k})$ and $\delta_m(\mathbf{k})$:

$$\delta_F(\mathbf{k}) = b(1+\beta\mu^2)\delta_m(\mathbf{k}) + \epsilon \tag{3}$$

$\epsilon \Rightarrow$ noise.

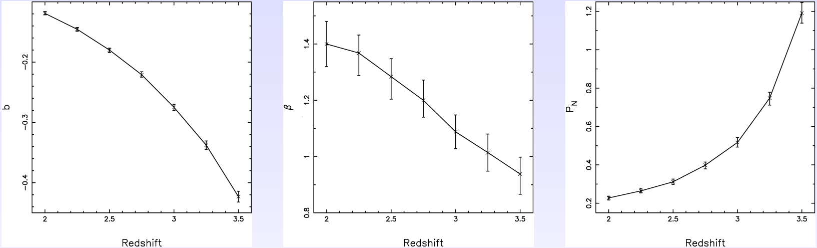
- Assume that ϵ is a gaussian random variable with variance:

$$\langle \epsilon\epsilon \rangle = P_N \tag{4}$$

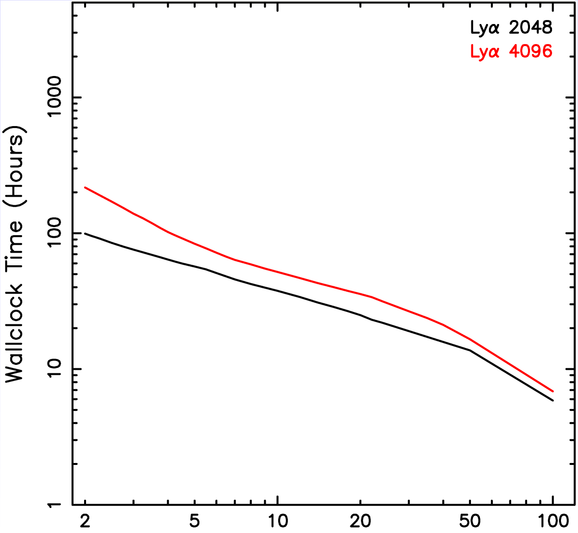
- Assume that ϵ is scale independent.
- One can then fit for b, β and P_N by minimizing:

$$-2\log\mathcal{L} = \sum_{i=1}^{N} \frac{\left[\delta_F(\mathbf{k}) - b\left(1+\beta\mu^2\right)\delta_m(\mathbf{k})\right]^2}{2P_N} - \frac{N}{2}\log P_N \tag{5}$$

The Evolution of Bias, β and Noise



Performance of the Lyα Forest Simulations



Proposed Runs

- Gadget3 scales very well on upto $\sim 10^5$ cores.
- We plan on looking at the dependence of the clustering of the Lyα forest on cosmological parameters.
- Running a grid of models for 4096^3 size simulations is expensive.
- Assuming that the scaling holds, a simulation with $L_{box} = 50$ Mpc/h and $N_{par} = 2 \times 896^3$ will take $\sim 500,000$ SUs.
- This estimate is conservative since there are fewer rare peaks in $L_{box} = 50$ Mpc/h as compared to $L_{box} = 400$ Mpc/h.
- First we need to establish resolution convergence and we are doing this now
- 2013 ERCAP proposal for 10^6 SUs for 20 cosmological models.

Code comparisons

- We plan to do code comparisons against Nyx - very different (SPH/AMR)
- Will make it easier to establish convergence of both codes
- Need to think about what to compare and when to call it an agreement
- Need to build a code-to-data pipeline to see how stable data fitting is wrt to underlying simulation technology
- We have just started this effort at this very workshop...